DOCUMENT RESUME

ED 219 397 TM 820 401

AUTHOR Bligh, Harold F.; And Others
TITLE Issues in the Development of an Individually-Administered Achievement Test.
PUB DATE 21 Mar 82
NOTE 41p.; Papers presented at the joint Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education (New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Criterion Referenced Tests; Data Analysis; Elementary School Mathematics; *Individual Testing; Norm Referenced Tests; Primary Education; Reading Achievement; Reading Comprehension; *Screening Tests; *Test Construction; Writing Evaluation
IDENTIFIERS Basic Achievement Skills Individual Screener; Psychological Corporation; *Rasch Model

ABSTRACT

Basic Achievement Skills Individual Screener (BASIS) was designed by the Psychological Corporation to yield criterion-referenced information as well as normative scores. The constraints imposed by this dual purpose were increased by the requirements inherent in an individually-administered format and the short testing time consistent with test use as a screener. Assessment of reading comprehension was determined to be the most valid and valuable aspect of reading to measure. The BASIS reading test employs a modified cloze technique. The mathematics portion was designed to be a quick screener yielding criterion- and norm-referenced information. The Writing Exercise is an optional subtest of BASIS still in the early stages of development. The Rasch model was used for data analysis because the BASIS tryout provided an opportunity to evaluate it using a small sample and a small number of items. Vertical equating would place all items and persons on a common scale. It would provide a method of validating grade placement of objectives being measured. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219397

Issues in the Development of an Individually-Administered Achievement Test

Joan L. Sonnenschein

Organizer

The Psychological Corporation

Harold F. Bligh

Chair

Participants

Eugene G. Bowles

Joan L. Sonnenschein

Lisa Cohen

Irene Swidler

The Psychological Corporation

Discussant

Gary J. Robertson

American Guidance Service

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Bowles, Swidler, Cohen, and Sonnenschein

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A symposium presented at a joint session during the annual meeting of the American Education Association and the National Council on Measurement in Education
New York, New York
March 21, 1982

TM 820 401

OVERVIEW

There seems to be currently two types of individually-administered achievement tests: either long and comprehensive for diagnostic use or short for screening purposes. These latter are exclusively norm-referenced instruments. Users, however, often attempt to interpret the resulting scores in a criterion-referenced fashion. That is, they use a grade equivalent score, for example, as though it represented the grade level of material the youngster is able to handle. They may then assign the youngster to a graded classroom or place him or her in a special program according to this grade equivalent. Such placement is often based on the apparent deficit between the observed grade score and the current or expected age/grade placement.

This is acutally a misuse of grade equivalent scores which have nothing to do with the acutal graded level of the material being tested, but rather represent the degree of agreement or departure from the score an average pupil at a given grade placement would obtain on this material. Put another way: The material may be entirely at a Grade 2 level; grade equivalents of 2.0, 3.0, 4.0 would then represent the performance (on THIS MATERIAL) of average students at the beginning or Grade 2, Grade 3, or Grade 4. I'm sure you're all aware of this, but I repeat it in order to show that -- again -- the level of the material was never a part of the grade equivalent concept.

Although many users are AWARE of the psychometric reasons underlying the inevitable inaccuracies in thus using grade equivalents as though they referred to the level of the material, they continue to use such a procedure for want of a better one, for want of more accurate information. Experience has led them to anticipate the degree of inacccuracy and to "adjust"

accordingly. You hear things like, "I know that score's always about 6 months too high, so I place 'em 6 months lower than the score indicates!" Such adjustments may be equally as inaccurate for the purpose as the original score, since there is no relationship between the purpose and the score.

In order to provide specifically the kind of information users have sought from existing tests, Basic Achievement Skills Individual Screener, (BASIS) was designed by The Psychological Corporation to yield criterion-referenced information, as well as normative scores. That is, the criterion-referenced information will indicate approximately the grade-referenced level of the material on which the youngster performs adequately.

This, plus the intention of deriving norms, placed a double burden on the material developed for the test. Furthermore, the constraints imposed by such a dual purpose were heightened by the requirements inherent in an individually-administered format and in the short testing times consistent with its use as a screener.

Such problems were common across the subject matter content areas included in the test. But there were also problems unique to each of these areas based on the requirements of the subjects themselves.

BASIS consists of three subtests -- Reading, Mathematics, and Spelling -- with an optional Writing Exercise. The three main tests are similar in these administrative procedures:

1. They are composed of clusters of items which are grade-referenced.
2. There are basal and ceiling rules so that only the most critical grade-referenced clusters are administered.
3. There are cutoff scores for each grade-referenced cluster to determine a final grade-referenced level of achievement for that content area of the test.

The first two speakers, Lisa Cohen and Irene Swidler, will describe more specifically the formats of the Reading and Mathematics Tests and will discuss the problems encountered in developing these two tests.

ASSESSING READING ACHIEVEMENT IN AN INDIVIDUALLY-ADMINISTERED SCREENER: CRITICAL ISSUES AND DEVELOPMENT

In developing a new individualized achievement screener, one of the first content decisions to be made was -- what area of reading will be tested? It is our opinion that reading comprehension is the essence of reading. We read to understand the author's message; we read for meaning. Therefore, a measure of a student's comprehension should yield the best estimate of a student's reading ability. Since BASIS is designed to be a screening device, and therefore, of necessity, a short test, we decided that the assessment of reading comprehension would be the most valid and valuable aspect of reading to measure.

We reviewed a number of format options before we determined a suitable method for assessing comprehension on BASIS. We decided not to go with the more traditional passage plus multiple choice questions because we felt that the individualized setting lent itself to utilizing a more individual or free response. We rejected the passage plus open-ended questions format (a style similar to many Informal Reading Inventories) for two reasons. First, the time factor was inhibiting. Second, the problem of determining if the difficulty level of the questions matched the difficulty level of the passage was an unnecessary additional constraint. Therefore, we chose to go with a format where the items are embedded in the passage and there is a free response style -- the cloze technique. Using an open-ended, modified cloze

format we were able to avoid the problems of multiple choice questions, economize on the time necessary to administer the reading test, and truly take advantage of the individual setting by requiring a free-response to the cloze items.

Research has shown that the cloze technique can be a very effective measure of comprehension. For instance, New York State adopted a cloze format for their test of reading competency -- The Degrees of Reading Power.

While reading a BASIS cloze passage, the reader's task is to fill in the blanks as s/he reads the selection aloud. The reader has to predict the missing word in the sentence by utilizing his/her understanding of the previous text. Comprehension can be defined as the ability to process information provided in the text to arrive at predictions which make sense within its context. Since the cloze task requires the reader to make predictions in order to complete a break, our reading passages are truly tapping the essence of the reading-comprehension process.

Strict cloze, generally used as a readability measure, makes deletions on a mechanical basis, usually every 5th or 10th word. BASIS uses a modified cloze technique where deletions are based on semantic and syntactic considerations. In order to achieve consistency in passage design, specific rules for writing our cloze passages were generated. We came across some interesting challenges while trying to apply these guidelines. As with most rules, it was fairly easy to apply one or two, but when there were nine or ten different restrictions they tended to confound one another and made writing the cloze items embedded into the passages a very challenging proposition. I would like to

explain the specifications that were set for writing the passages, the cloze items, the readability formula applied on top of these specifications, how successful we were in adhering to these constraints, and some of the methods we employed to help apply them.

To begin with, the first 20 words or two sentences in each passage were to be left intact to provide the students enough context to orient themselves to the passage. With very few exceptions, the passages met this requirement. Thereafter, one word was to be be deleted approximately every 20 words. We were fairly successful in our ability to adhere to this rule. Although every cloze break is not preceded by 20 words of running text, the mean number of words between cloze breaks is 21 and, at the lower limit, not less than 9 words, which represents at least one complete sentence, between items. Next we aimed to have every cloze break appear at the end of a sentence or phrase. This would allow the reading to flow and would not require the student to have to read beyond the break in a sentence to be able to complete the cloze. For the final form, in 85% of the items, the break is the last or next to last word of the sentence. So again, we were able to apply this rule successfully. The next constraint dealt with the type of words to be deleted. Since we were aiming to assess a student's comprehension of some portion of continuous text, we decided it was best to delete only content words; that is, nouns, verbs, adjectives and adverbs. In addition, we tried to have a fairly even distribution of these types of deletions across each passage. Again, we were able to meet this requirement.

Along with specifying the part of speech, the word to be deleted had to be at a grade level below that of the passage. Furthermore, the majority of deletions were to be at the pre-primer, primer or 1st grade level. We

verified a word's grade placement in the Harris-Jacobson Basic Elementary Reading Vocabularies. We placed this constraint on the items because we wanted to avoid the possibility of a student's poor verbal fluency unfairly hampering his/her chance to receive credit for completing a cloze break. We have successfully applied this rule to the cloze items. Every passage's cloze breaks have exact restorations which are at least 1 grade level below that of the passage, and they are usually more than 2 or 3 grades below at the upper levels. In fact, no restorations is above the 5th grade level. According to our readability formula, we allowed some words in the passage to be above grade level, but the cloze item could not be dependent on the understanding of the above level words. Also, it was preferable that a word used as an item in a passage did not appear in the previous text of the passage. We found that this was not always possible to achieve at the lower levels. Due to the limitations of the primer and first grade vocabularies it was very difficult not to have one item that repeated a word in the passage. The vocabulary is just too restricted to come up with meaningful passages and items otherwise.

Now we come to the most important constraints placed on building the items. The cloze items had to be independent of one another. They could not deal with the same concept or information within the passage. It was our goal that responding correctly to one item should not be contingent upon having answered a previous item correctly. In addition, the item had to be dependent on the text. It could not be answered unless the preceding text was understood. Also, the sentence containing the cloze item could not be self-sufficient; if taken out of context it could not be answered correctly. Responding correctly depended on the reader's understanding of the previous sentence, sentences, or

paragraph. Please note that the correct answer could not be dependent on text that follows the item. We did not allow the students to return to an item after having continued reading. Only spontaneous corrections were allowed.

In order to help assure that our items were indeed text dependent, we conducted a small informal study. For all the items in all the passages we were considering (there were approximately 5 or 6 passages per grade level), we took all the sentences which contained the cloze breaks out of context and listed them randomly, one page per grade. Then we asked 10 people from the Measurement Division at The Psychological Corporation to fill in the blanks to the best of their ability. When we compiled the results, we found that in some instances half of the respondents could come up with a correct response without reading the passage! We then revised the items found to be problems and repeated the process. We continued revising the items until we found that the items could not be answered correctly with any consistency. Of course this is in no way a scientific study, but it did help us focus on the factors that helped the readers fill in the blanks successfully and therefore enabled us to rewrite the items with special considerations in mind.

In addition to all these restraints, the passages were carefully controlled for readability. A similar procedure as that used successfully on the Metropolitan Achievement Tests, 1978 edition, was employed. The Harris-Jacobson Basic Elementary Reading Vocabularies was used for grades Primer to 6 and the EDL CORE Vocabularies was used for grades 7 and 8. The CORE Vocabularies word lists were modified to eliminate any words that appeared at lower levels in Harris-Jacobson. For grades 1-6, a passage had to contain a minimum of 25-30% on level words. For grades 7 and 8, passages

contained 20-25% on-level words. At primer, the passages contained approximately 90% on-level words. For grades 1-8, passages could not contain more than 5% above-level words and at primer not more than 10%. At every grade level, no word could be used that was more than 2 grades above level.

For Item Analysis, each primer passage contained two cloze breaks and was approximately 50 words long. First grade passages contained three cloze breaks and were approximately 75 words long. Passage in grades 2 through 8 contained seven cloze breaks and were approximately 150 words long. In the final form, there will be 3 primer passages with 2 cloze breaks each, 2 grade 1 passages with 3 cloze breaks each, and 1 passage each in grade 2 to 8 with six cloze breaks apiece. Therefore, in the final edition there will be six cloze items per grade level.

Considering all the restraints of the modified cloze format, on top of the readability formula, we recognized that the sentences with the cloze breaks would necessarily be fairly short, contain very little information within themselves, and taken out of context would tend to be ambiguous. Therefore, we anticipated that students who arrived at correct responses based on the previous text, would answer the items with a variety of synonymous responses. The directions to the test examiners stated we were not restricting an acceptable response to the exact restoration, an acceptable response must have essentially the same meaning as the original restoration, and it must make sense within the context of the passage.

When the scored passages were returned and compiled, however, we found that the ambiguity of the sentences lead to responses of a wider variety than had been anticipated. Not only did we find synonymous responses, but there were also many instances of nonsynonymous but potentially correct responses. There

appears to be two reasons for this. First, because the task was open-ended, and because they could not necessarily tell the direction in which the story would later go, the readers were able to emphasize aspects of the text that were different from our intention when writing the items. Therefore, their responses were in no way synonymous to the exact restoration, but they did make sense within the context of the passage and could not be considered incorrect. Second, some readers seemed to use their outside experience when responding, rather than relying on the information in the text. In most cases, though these responses made sense in the sentence itself, they could not be substantiated in the text and were therefore considered unacceptable. When we proceeded to check score the student protocols we quickly learned that the examiners had scored the responses using their own background and clinical judgment which often yielded conflicting results. Since it appeared that the open-ended response mode produced responses on a continuum of acceptability which had to be refined into correct and incorrect categories, and since it was clear that the test examiners had not consistently agreed on scoring these responses, we revised our scoring procedure.

First, we clarified our criteria for scoring cloze breaks. It was decided that the response should be both semantically and syntactically correct. Semantically correct was defined as (a) either a synonym or closely associated word, or (b) a word predicted from the context of the passage and not from outside experience. Even if the reader's experience leads him/her to say a word that could be true in the passage, but is no way stated or implied in the passage, it is not an acceptable response. We have to assume that the answer, though possibly triggered by some information in the text, came from the reader's outside experience and not from a full understanding of the text.

Next, we reviewed all the responses and coded them with either correct, incorrect, or special codes for those responses which had a great inconsistenccy in scoring and were on the borderline of acceptability. Then we generated a frequency distribution broken down by grade, sex, and race. In this way we were able to determine if any borderline responses were found more commonly among any specific group and the extent to which a specially coded response was given. After reviewing this data we made our final determinations as to which special codes were to be considered correct and which were to be considered incorrect. It should be noted that in some instances we considered some syntactically incorrect responses acceptable if they showed a good understanding of the passage. Then all the protocols were rescored according to these guidelines to assure consistent scoring.

Once all the data was in, the last major decision was determining which passage would be placed in the final form. Our item analysis design insured that there would be 2/3 more passages tried out than would be used in the final format. Also, in grades 2-8, there would be one extra cloze item in each passage which would be restored.

First, a decision was made as to which six items to retain in each passage. This was done by analyzing the individual item p-values, point biserials, and the average p-value for the different possibilities of a six-item passage. We checked the range of p-values and the possibility of difficult scoring of each item. Since each passage was read by students one grade level above and below the designated grade level, in addition to the students in the targeted grade, one of the most important considerations was evidence of growth in the difficulty level of the items, particularly between the below-grade level

group and the on-grade level group. Once one of the seven items was discarded in each passage, it was also possible to eliminate a few passages from further consideration. We were then able to make a decision as to which passage to retain at each grade level. Based on the knowledge and experience of our resident experts, we predetermined the criteria for an appropriate passage at each grade level. It was decided that approximately 70-80% of on-grade students taking the passage (late in the school year) should pass, in order for the passage to be considered acceptable for that grade level. Furthermore, the performance of the below-grade level students should show a marked decline in ability to pass the passage.

In order to find the best passage at each level, we had the data analysis rerun for six-item passages and generated empirical frequency distributions. These frequency distributions enabled us to determine how the items performed on a group basis, rather than on an individual basis. As stated previously, it was this group performance that was looked upon as the key factor in choosing the final passages. With these frequency distributions we were able to investigate the results of using various cut-scores with each passage in order to find the passages that would best meet our criteria. We found that in order to maintain our pre-set passing criteria of 70-80% of the on-grade students, it was necessary to vary the cut scores at different grade levels. Although it would have appeared neater and easier to score had all the cut scores been the same, we concluded that this would not be possible if we were going to maintain a criterion-referenced test. It should also be noted that although our cut-scores allow students to pass a passage with scores of between 50-67% correct, this is considered a very acceptable level of accuracy for a cloze test. Studies have shown that the traditional 75%-90% passing range on multiple choice tests is equivalent to the 40-60% range on cloze

tests. So, even though the passing percentage may appear low to those familiar with the conventional 75% pass level, taking the cloze format into consideration, it can be shown that the BASIS reading passages have been designed to indicate a student's instructional reading ability.

In its final form, the reading section of BASIS will contain passages from Primer to Grade 8 -- 3 Primer, 2 Grade 1, and 1 each in Grades 2-8. The directions to the examiners will include the lists of acceptable and unacceptable responses we compilated during item analysis. In this way we will insure that each student's protocol will be easily scored in a consistent and reliable manner. In theory, it is likely that each student will read three passages before reaching ceiling. This will enable the examiner to determine the student's instructional reading grade which is the essential criterion-referenced information to be gained from the reading assessment section of BASIS.

Measuring Achievement in a Mathematics Screener -- Time Vs. Coverage

The mathematics portion of BASIS was designed to be a quick screener that would yield both criterion-referenced and norm-referenced information. To keep testing time minimal, and yet provide users with both types of information, it was necessary to determine which objectives were most important at each grade so as to assess the "essence" of achievement at that grade in the fewest possible items.

First, research was undertaken to determine which objectives were typical of a given grade's curriculum. This included a review of 10 major math textbook series, six math achievement tests, summaries of national teacher questionnaires on their classrooms' math curriculum, state mathematics manuals, various related articles, an informal survey of teachers, and personal teaching experience.

It was obvious that there were too many different areas of mathematics (concepts, numeration, geometry, measurement, graphs, statistics, algebra, computation, and problem solving) included in each grade's curriculum to allow a selection of a representative sample from each domain, given the time constraints of a screener. The areas at the core of mathematics, however, which are most stressed in the classroom, are computation and problem solving. Computation has long been the heart of mathematics, and so would be emphasized on BASIS. The importance of problem solving is currently gaining widespread recognition. A recent recommendation of the National Council of Teachers of Mathematics was: "Problem solving must be the focus of school mathematics in the 1980's."* Thus, we decided to limit the math portion of BASIS to these two domains.

*The National Council of Teachers of Mathematics, Inc., An Agenda for Action,

Within the areas of computation and problem solving, our research revealed that it is often impossible to pinpoint an exact grade level for many specific skills. The content of all the tests overlaps somewhat; textbooks vary in the grade levels at which a skill is introduced, emphasized, and reinforced; teachers of the same grade level teach some of the same things at the same time, but stop at different points. Yet, we found a definite similarity for many objectives as to when they are predominantly taught, and a general concurrence on the hierarchy of skills in the four operations from grade to grade. By selecting only those objectives for which there was the greatest agreement about grade placement, a core group of 9-11 objectives for each grade, from Readiness level through Grade 8, was determined. At this point, sets of items were written for each objective. Two items per objective were chosen for tryout, one for each of two forms, based on comparable estimated difficulty, and a balanced digit and number fact representation.

We strove to make the word problems clear and realistic, with a vocabulary control so that the readability of each word problem was on or below its assigned grade level. This precaution was taken in order to maximize comprehension, despite the fact that all word problems are dictated on BASIS. Students are also free to read and reread the word problems, so that both aural and visual comprehension can be utilized. The word problems have positive ethnic representation and equal portrayal of both sexes.

Research on typeface and typesize included consulting such organizations as: The New York ORTEN Society, The National Association for the Visually Handicapped, The Educators Publishing Service, Inc., The Institute for the Study of the Learning Disabled, and the Council for Exeptional Children.

Information was obtained on appropriate leading, density, clarity, and degree of familiarity of various typefaces. This research, combined with a survey of math workbooks and textbooks, enabled us to select the clearest typeface and the most appropriate type sizes for the computation problems at each grade level. Both clarity of form and size were considered important to minimize the perceptual problems some students have. The appearance of the items and the open-ended format were chosen to make the test items most similar to the regular classroom work with which students are most familiar and comfortable.

We tried out these items on students who had scored in the 4th - 6th stanine range on a previous math achievement test. Therefore, they could be considered "average-achieving" math students. Each cluster of grade-referenced items was given to students in the actual grade at which the objectives are predominantly taught and to students in adjacent grades.

The data from item tryout included *p* values and point biserials for each item for below-grade, on-grade, and above-grade pupils. We wanted to select a group of items for each grade level that would be not only psychometrically sound but that would reflect the core of that grade's curriculum.

The data revealed that the range of item difficulty for students at each grade was quite wide (at least .60 at all grade levels except Readiness.) Thus, the items ranged from quite easy (*p* values in the .90's), to very difficult (*p* values in the .30's or below). Of the 178 math items tried out, 70% had *p* values greater than .50 for on-grade pupils.

At this point, we strove to include in each grade's cluster of items, one or two objectives for each of the basic operations which are part of that grade's scope, so as to present a balanced representation of the variety of skills that are acquired each year. This hierarchical progression of skills in each of the four basic operations would indicate to the test user how far a student has progressed in skill acquisition for a particular operation, providing a point to begin further diagnostic exploration, if needed.This can be especially useful in some cases of learning disability, where growth in the four operations is often uneven.

To keep testing time minimal, and yet adequately cover each grade's core curriculum, we decided to form clusters of 6 computation and 2 problem solving items at each grade level (except for the Readiness level), with each item assessing a distinct objective. Some objectives tried out had to be eliminated because they were too easy (eg. multiplication of basic facts at Grade 3), or too difficult (eg. subtraction of decimals in horizontal format at Grade 6) at their appropriate grade placement. We could not place these objectives at other grades based on their psychometric characteristics, because they would not be curricularly appropriate there.

In building a test that would be both criterion and norm-referenced, we could exclude the extremes of difficulty only if we could still obtain a core group of representative items for each grade. The items selected for Grades 1-7 were not only representative of their grade's curriculum, but had a difficulty range of .50-.85. At the Readiness level, we included some easy items (p values in the .90's) since testing was conducted in the spring. Our on-grade sample for this level consisted of below-average first graders although our results show that these students learn Readiness level material

by the spring of Grade 1. We expect that below-average beginning first graders, tested in the fall, will not find these items quite so easy and will be correctly placed on Readiness material. On the other hand, we selected two difficult items (p values in the .40's) for our Grade 8 cluster to reflect the diversity of the curriculum at that grade.

In our final selection of items, the items generally became gradually easier across the 3 or 4 grades at which they were tried out. The fact that this increase in item easiness was usually gradual rather than sudden may be due to the overlap in instruction over successive grades. No item selected for the final test showed a significant reversal in terms of item difficulty between grades.

For slightly more than 10% of the items, however, there was a great amount of growth in the success rate between students below the grade where that objective is usually taught and students at that grade level. In some cases, the difference in p-values between the below-grade and on-grade samples was greater than .40. For example, the multiplication problem, 37 x 5, yielded p values of .29 for third graders, and .88 for fourth graders. This phenomenon demonstrates that some skills are clearly acquired at a particular grade level. In this instance, the skill of multiplying two-digits by one digit with regrouping once is usually introduced at Grade 3, but it is not until Grade 4 that all texts, and presumably all teachers, have taught this skill. This objective was selected for Grade 4 on BASIS since it is emphasized in the Grade 4 curriculum, as an important part of that grade's learning, and because it discriminated so well between 3rd and 4th grade students.

Another important aspect of the data analysis was to avoid a type of item difficulty reversal that is best explained by example: A fourth-grade student fails to pass the fourth-grade cluster of items, and is therefore not permitted to attempt any items in the fifth-grade cluster. If the fifth-grade cluster were to contain any items that were easier (for fourth-graders) than the hardest item in the fourth-grade cluster, he would not be given the opportunity to attempt an item that he might have been able to answer correctly. In the very few instances where such absolute reversals could not be avoided, their impact was neutralized by the minimal amount of the overlap. That is, although the grade-five item might be easier by only one or two p value points, it would not have been possible to fail the fourth-grade cluster by answering incorrectly only the one harder grade-four item involved in the reversal. (He would have had to answer one or two others incorrectly as well.) Nor does this minimal reversal affect the pass/fail potential outcome of the fifth-grade cluster.

Next, we sought to determine how well each grade-referenced cluster performed as a group, Ideally, one would expect an average pupil to get almost all of the items right targeted to the grade level at which he or she is functioning, all of the previous grades' items right, and none of the successive grades' items; but few average pupils match this ideal performance. We decided that our clusters would be functioning acceptably if at least 70% of the on-grade pupils could pass most of the material in their grade-referenced cluster. Thus, we expected that 5 or 6 out of the 8 items in each grade-referenced cluster should be answered correctly by at least 70% of the on-grade students.

In math, because of the two forms in our item try-out design, no student took all of the items in a given cluster selected for the final test. As such, we could not generate an observed frequency distribution, as was done with Reading. Instead, we generated expected frequency distributions, using the binomial expansion formula, to yield expected percentages of pupils passing at each grade at various cut-scores. This process indicated that from 72-80% of the on-grade pupils could be expected to answer correctly 5 out of 8 items at their grade level, for Grades 1 - 7.

For the easier Readiness cluster, we chose a more stringent cut score. We found that 80% of our below average first graders could be expected to pass the Readiness level with a cut score of 5 out of 6 items.

Because the Grade 8 items were more difficult than those at the other grades, the cut-score for passing eighth grade math was lowered to 4 out of 8, at which point 77% of the 8th graders passed.

Of course, these cut scores will be reevaluated based on the empirical data from the standardization program.

We used the same procedure and the same cut-scores to determine the expected percentage of below and above grade pupils passing each grade-referenced cluster. We found that the expected percentage of below-grade pupils passing the next grade's cluster was between 3%-40% across grades. At Grades 2 - 6, the percentage of below grade pupils passing the next grade's cluster was 15%, or less. The probabilities of above-grade pupils passing the previous

grade's cluster was 87% or greater. At the Readiness level through Grade 5, the percentage of above-grade pupils who could be expected to pass the previous grade's cluster was 94% or more. These results were clearly in line with our goals of having distinct grade clusters that get progressively more difficult, and discriminate between pupils at successive grade levels.

In conclusion, the item tryout for the math portion of BASIS confirmed most of our preliminary research regarding the grade-placement of objectives. There were sufficient items that most on-grade students answered correctly at each grade level, to allow us to select a balanced cluster of items representing each of the operations stressed at that grade level. We have included items assessing the hierarchical progression of skills within the four basic operations from grade to grade. We achieved our aim on constructing distinct grade-keyed clusters, each reflecting the essence of that grade's curriculum and discriminating well between pupils at adjacent grade levels.

Issues in the Development of an Individually-Administered Writing Exercise

Classroom teachers and entire school systems have become increasingly interested in assessing the writing skills of their pupils. The use of a work-sample technique, rather than a multiple-choice test which measures knowledge of writing mechanics, has gained in popularity over the last few years.

There are, however, problems inherent in the use of a writing sample test. The first difficulty is one of standardization so that comparisons among written outcomes can be made. Such standardization necessitates that all the writers be addressing the same topic, all writing in the same mode -- descriptive, narrative, etc., -- and all adhering to the same time limits. This presupposes that they will all be stimulated readily and fairly equally by the intriguing topic chosen and will all have their creative juices turned on simultaneously!

The second difficulty has to do with scoring or evaluating the papers. Some subjectivity must inevitably be involved and good reliability is not easy to achieve.

A third difficulty is certainly the fact that this methodology is quite time-consuming and therefore costly.

It is an interesting aside that many of the recently developed programs and guidelines for the teaching of writing skills pay scant attention to the assessment of the efficacy of their proposed techniques.

The two scoring methodologies usually employed in writing assessment are part scoring and holistic scoring. In part scoring the scorer is asked to address various facets of the writing independent from one another. These facets may include such things as spelling, punctuation and capitalization, richness of vocabulary, variety of sentence structure, organization, and, perhaps, creativity. Often, the scorer is asked to review the paper specifically for one facet and then review it again with a second facet in mind, etc. Scores are then assigned for each of the facets of interest. This may require quite a bit of time per paper. Also, there must be carefully-worded guidelines indicating, for each facet, what can be expected of students at the given grade.

Holistic scoring is the more popular method. In this technique the scorer reads the paper only once for an all over general impression. That impression is then matched against a set of previously scaled papers in order to assign a rating. The scaling of the set of initial papers which make up the standards, and the matching of each new paper consistently to the same point, and to that point that would be selected by other reviewers, all require training and practice among the reviewers.

Furthermore, the number of scaled points is usually set before the papers themselves have been addressed. That is, if this is to be an 8-point scale,, for example, with a rating of 8 representing the best papers and a rating of 1 representing the least worthy papers, such a decision is made before the range of quality in the given set of papers has been ascertained. This means that, since the papers selected to represent the 8 points on the scale come from the batch of papers to be scored, the best of this batch will get the top rating

and the worst of this batch will get the lowest rating -- regardless of the absolute quality of the writing in these papers. These papers are usually limited to those written by students at a single grade level or a narrow range of grades.

This holistic technique presupposes that a group of papers -- from a single classroom's worth to the output of an entire school system -- are to be compared, and that the reviewer is trained and experienced in holistic scoring and is probably familiar in general with the range of quality of writing that is usually produced by students at this grade.

When we started to plan the Writing Exercise that would be an optional test in Basic Achievement Skills Individual Screener (BASIS), we realized that virtually none of these scoring conditions were likely to prevail. That is, since this is an individually-administered test, the writing reviewer would be dealing with a single paper at a time, rather than a group; succeeding testees were likely to be at different grades or, perhaps, in ungraded situations, so that the reviewer probably could not reasonably wait to accumulate same-grade papers for comparisons -- not to mention that, as part of a screener, one would expect and require rather immediate feedback; and finally, the test administrator was more likely to be a school psychologist or a resource room teacher than to be a trained holistic scorer -- or indeed a current regular classroom teacher who would be completely familiar with the writing skills to be expected of a student at any specific grade level.

It was therefore decided that we could require the test administrator to make no finer distinction or rating than whether a given paper seemed to match, in a general way, an "average" paper of the same grade or age; or, if it did not, whether it seemed to be "above" or better than the "average" paper or worse than and therefore "below" that "average" paper. It was obvious that we would have to supply that "average" paper or standard in the test manual.

This also suggested that the amount of written material produced should be reasonably short -- to match the time expectations of a screener and to facilitate the reading and matching of the papers and the standards. The topic selected therefore had to be simple and familiar and one that would elicit descriptive writing -- which research has indicated is easiest for grade school children to handle. The topic chosen was "My Favorite Place."

To field test this topic and to derive the "average" papers needed as scoring standards, 2 school systems -- one in Michigan and one in Florida -- of average socioeconomic levels, administered the writing sample, in the closing months of last year, to one of their typical classrooms at each of the target grades, 3 through 8; a total of about 60 papers per grade. In order to stimulate the writers and to focus their writing, the following directions and guidelines were given after a few, brief introductory remarks:

> Write about your favorite place. It could be a room, a house, a park, or any other place that is special to you. Describe the place in any way you like, but be sure to include what makes this your favorite place.

After time for questions and a repetition of the topic, they were told, "You will have ten minutes to write about the topic." At the end of 8 minutes they were warned, "There are 2 minutes left. You should be finishing your papers." At the end of ten minutes, the teacher said, "Time is up. You may finish the sentence you are writing and then put your pencils down."

Copies of these papers were then mailed, in January of this year, to 6 trained, experienced holistic scorers from three different parts of the country, who were affiliated with either a recognized writing program (such as the Bay Area Writing Project) or a writing research program (such as the one out of Miami University in Ohio). The scorers were asked to score holisticcally according to the following instructions:

1. For each grade level, score each paper on a three-point scale -- below average, average, and above average.
2. Re-evaluate the average pile to determine the nine papers that are the average of the average.
3. Rank those nine papers 1 to 9, with 9 the best.

In adddition, after completing all scoring, the scorers were asked to list some common characteristic or salient features that were indictative of average papers at each grade, using the complete pile of average papers. They were asked to focus on the presence or absence of these various characteristics and their significance at the given grade; the differences, at each grade, between average papers and those labeled either above-average or below-average; and finally, to note developmental patterns, if any, across grade levels.

As of this date, we have not concluded statistical analyses of the concurrence among these raters -- the papers were just returned recently -- but it is obvious that there is, in general, a great deal of agreement, so that we will have little difficulty in determining an "average" paper to be used as the standard at each grade. For example, at most grades, at least four of the six raters assigned a number to the same paper, while the remaining raters had marked it as average but unnumbered.

We have yet to determine whether the lists of characterisitcs agree and whether it will be possible to suggest clearcut distinctions between average papers at adjacent grades.

It must be pointed out that, so far, no one has actually attempted to match any one paper to a preselected average paper for a given grade -- which is the task intended for the final test. We have thus far been establishing only the following important points:

1. The topic and directions do elicit writing samples that are sufficiently detailed and complex to allow discrimination among papers at each grade.
2. Students in the middle of Grade 3 can handle the task.
3. The 10-minute time limit does allow even those with a good deal to say to express themselves fully.
4. The samples thus produced can be rated holistically on a three-point scale.
5. An "average" paper can be identified at each grade with a high degree of agreement among trained, experienced raters.

The scoring procedure itself will be validated during the standardization of BASIS which is scheduled for this fall. At that time, test administrators -- usually graduate students with some training in the individual administration of tests -- will be asked to judge each writing sample by comparing it with the given "average" paper at the grade. He or she will determine whether the particular paper, in general, contains the characteristics of the average paper, lacks them and is poorly executed (a below-average paper), or contains them to a high degree and is well executed (an above-average paper). Some subsample of these papers will also be scored by trained scorers to determine a correlation between the trained scorers and the test administrator.

There is still a long way to go in the development of the Writing Test for BASIS, but we are encouraged to think that we will be able to provide a quick assessment of a student's writing level. We believe that that assessment, when added to information about achievement in other academic areas, will prove valuable for general screening of the student's strengths and weaknesses.

In closing, I might add that the particular content of each of these writing samples, made interesting reading from a counseling perspective.

Psychometric and Research Issues in the Development of an Individually-Administered Research Screener

Overview

The BASIS Item Analysis Program was conducted during the spring of 1981. The tryout material was presented in 2 forms (X and Y) each of 2 separate tests, Reading and Mathematics/Spelling. Each of approximately 3900 students in grades 1-8 were individually administered 1 form of 1 test. Each student took items that were on their grade-placement level in a curricular sense as well as items placed at adjacent grade levels.

* * *

In the initial planning stages of BASIS, the aim was to develop a test such that grade-referenced clusters of items would be the basic unit of test administration. The items in each cluster would measure those objectives that are commonly taught and presumably learned at a specific grade level. The intent for the final use of the test was to provide information concerning the appropriate grade placement of the student, from both a criterion-/and norm-referenced perspective.

As such, we were faced with the following considerations:

1. Defining the sample.
2. Testing a sufficient number of students to yield stable item statistics while also trying out each item at a sufficient number of grade levels to evaluate the grade placcement of the item.

3. Arranging to have a nationally representative sample of students individually tested.
4. Developing a research design that would allow us to validate the grade placement of each item.

The sample used in the BASIS tryout had to provide us with both valid item data and validate our grade placement of each objective being measured. This latter purpose was essential to our task since the test blueprint included the restriction that an objective should appear in the test only at the grade level at which that objective is taught and emphasized. Thus, even though an item measuring a Grade 5 objective was too easy at Grade 5 but was psychometrically sound at Grade 4, it would not be used in the Grade 4 cluster of items on the final test. Similarly, a Grade 5 objective that is quite difficult at Grade 5 but that is learned adequately at the later grades also would not be included in a post-Grade 5 item cluster.

A second consideration in defining the item analysis sample was the cost -- in both money and time -- inherent in individual testing. It was important to maximize the amount of information collected on a well-defined sample of students while minimizing both the number of students tested and the test administration time.

Therefore, to assure obtaining a sample of students with reliably defined achievement characteristics without adding additional testing time, we limited the sample to only those students who had scored in the mid-three stanine levels on a recently administered national or state-wide standardized

achievement test. An exception was an additional sample of Grade 1 students scoring in the lower third stanine range, designated Grade 0. For any student for which a standardized test score was not available, the classroom teacher's judgment was used.

The rationale for restricting the range of previous achievement performance of the sample was threefold. First, it would increase the probability of selecting homogeneous, small samples across grade levels. Second, we were concerned that the test examiners would encounter resistance on the part of schools and parents if a student interview were used to collect demographic data. Further, the accuracy of such data reported by grade school students would be questionable, at best. The use of previous stanine scores, on the other hand, would give us a standardized metric to evaluate both the success of the sample selection procedure and the homogeneity of the samples at each grade. Essentially, we assumed that a stanine 5 student was a stanine 5 student, regardless of any other demographic characteristic. Third, restricting the range of achievement levels to include only "typical" or "average" students would help identify those objectives that were most appropriate at each grade.

In effect, we attempted to guarantee that each grade sample would have a mean and modal previous achievement level of stanine 5. We assumed that a small homogeneous sample would yield more stable p-values than would a small heterogeneous sample but that the obtained point-biserial correlations would be conservative estimates. Further, since most students were to be tested on

content pegged both above and below their present grade placement, the assumption was made that these out-of-level data would provide sufficient estimates for non-average students, that is, students not in the 4-5-6 stanine range.

The final sample in Grades 1-8 was composed of approximately 57% and 61% stanine 5 students in the Mathematics/Spelling and Reading samples, respectively. For both samples, there were more stanine 6 students than stanine 4 students by about a 2 to 1 ratio.

The research design required 400 students per grade, Grades 0-7 and 800 students in Grade 8. At each grade, there were effectively 4 samples: 2 Reading samples (Form X and Form Y) and 2 Mathematics/Spelling samples (Form X and Form Y). Students were randomly assigned to one form of one test. Table 1 presents the expected sample size taking each grade referenced-cluster of items on one form of one test.

Insert Table 1

about here

It can be seen in Table 1 that each cluster of items, with the exception of pre-Grade 1 items, were tried out on at least 300 students. Within each cluster of items, certain items were overlapped across both forms, thus doubling the number of students that took them.

Of the 4000 students expected to participate in the program, 3918 were actually tested -- 1946 in Reading and 1972 in Mathematics/Spelling.

To include a nationwide sample of students in the item analysis program, and to insure that all students were tested at approximately the same time in the school year, 68 Field Supervisors were recruited from the faculties of graduate schools of education and psychology across the country. The role of the Field Supervisor was to coordinate the item tryout program at the local level. The major responsibilities of the Field Supervisors were to invite local school systems to participate, to recruit test examiners, to obtain necessary school and parental permissions to test selected students, and to monitor the actual test administration. The test examiners recruited by the Field Supervisor were mostly graduate students or local school system personnel with individual testing experience.

Although 13 of the original 68 Field Supervisors did not remain in the program through completion, the final sample included students from more than 90 school systems in 31 states. The sample included approximately equal proportions of males and females; and 85% White and Hispanic students, 13% Black students and 2% in the "other" category.

Most Field Supervisors had no trouble with testing their assigned quota of cases and many were able to test more at our request. A difficulty encountered by many Field Supervisors that they found surprising, to say the least, was the reluctance or outright refusal of a local system, in which the supervisor had a good, long-standing personal or working relationship, to participate in this type of research program. Whenever possible, we asked the

Field Supervisor to find out specifically what it was about this particular program that the school objected to. In many cases, the schools felt that, given their own testing programs and additional state mandated testing, their students were already being overtested. This is, of course, a legitimate and understandable reason. In other cases, however, the objections were more difficult to deal with. First, some schools did not want to participate in any type of research program. Second, the fact that this test was to be administered individually and that previous test scores were needed posed a problem. When probed, a number of schools indicated that they didn't want to go through the school board and/or parental approval procedure process, or deal with the apparent confidentiality issue. A number of Field Supervisors, especially those teaching graduate level courses in school administration, indicated that they were going to devote more class time to the proper evaluation of research proposals and to the role of the school system in test development. It will be interesting to follow this up and see what their graduate students' reactions are to this issue. We are presently planning 2 surveys on this topic, one directed to school administrators and the other to the faculty of graduate education departments. We plan to present the results at a future meeting.

The research design was developed to try out all items at the on-grade level and at grades adjacent to the target grade, to keep testing time at a reasonable length, and to attempt vertical scaling of the test items using the Rasch model.

To meet the first two goals, each student in the Reading Test sample was administered 6 grade-referenced clusters, in this case, reading passages, -- 2 on-grade and 2 from each adjacent grade level. The Mathematics/Spelling sample was similarly tested on 3 grade-referenced clusters in each content area. As stated earlier, this resulted in approximately 300 students taking each cluster of items. The average time needed to administer the test was about 45 minutes.

In addition to providing an index of growth, the testing of items at adjacent grade levels also provided the necessary links for Rasch equating of levels. To equate forms at the same level, certain items appeared on both test forms. In Reading, one passage per grade level was common to both forms. In Mathematics, 4 of the 12 items in a grade-referenced cluster were overlapped across forms. In Spelling, 2 of the 7 items appearing on each form were overlapped.

The purposes of using the Rasch model in the data analysis were twofold. First, the BASIS item tryout provided an opportunity to evaluate the appropriateness of the Rasch model using both a small sample and a small number of items. In our recent revision of the Stanford Achievement Test, the Rasch model was used to analyze over 35,000 items appearing in 146 booklets administered to more than 100,000 students. Since BASIS is essentially at the other extreme of the test type continuum, testing the Rasch model in this situation seemed worthwhile. Secondly, the vertical equating would place all items and persons on a common scale. This would be one method of validating our grade placement of the objectives being measured as well as identifying clusters of appropriate items in terms of item difficulty for the ability level of the grade to which they are referenced.

The data analysis was originally designed in a two-step process. The first step would yield both classical and Rasch item statistics, separate by grade. Geometrically, this step is represented by vertical slices of the design presented in Table 1. The second stage was to equate vertically all items appearing on the various experimental test forms.

Due to the problems that arose with the examiner scoring of the Reading Test, the data analysis was begun on the Mathematics and Spelling Tests while the Reading was being rescored. The analysis of the data generated in this first stage presented two problems. In Spelling, even though the words were carefully selected for tryout to be appropriate at a given grade level based on Spelling lists and a review of the major publisher's Spelling texts, the resulting score distributions were negatively skewed and most items would not calibrate. At many grades, the Rasch analysis would terminate at the editing process because too many items and persons were deleted from the data base. Basically, many of the words were too easy. A second analysis was conducted, separate by grade, using only on-level and above-level words. This analysis also resulted in too many items that would not calibrate.

The Mathematics item statistics were as expected for the most part, yet there was a certain amount of inconsistency in the Rasch item-fit statistic for particular items from grade to grade. In many cases, the observed inconsistency could be explained by the nature of the objective being measured. In some cases, the particular objective was apparently well-learned at the grade it was introduced in the curriculum and was then too easy for

students beyond that level. In this situation, the item would calibrate in the analysis for the on-grade students, but there would be essentially a content/sample mismatch at the grade above. The item failed to discriminate for the above-level students and thus, did not fit the Rasch model. A similar situation existed for other items that were too difficult at the grade level they are introduced in the curriculum, but were apparently learned at a later grade.

We then re-analyzed the Mathematics data, separate by grade-referenced item cluster. Geometrically, this would be a horizontal slice of Table 1. All items did calibrate using this approach and preliminary comparisons of the two approaches to the analysis suggested that the Rasch model was indeed appropriate. However, we were beginning to face serious time problems and developmental deadlines so it was decided to use the classical item statistics to select items for the final test.

After the decision was made to use the classical item statistics, the item selection was fairly routine. Many of the considerations that went into this process have been discussed already. The one aspect that was somewhat different than normal was the use of the binomial model to estimate expected frequency distributions for the final set of items that no one student had actually taken. Since the grade-referenced cluster of items is the basic unit of test administration, it was important to estimate how the final configurations of the items would act as a group. Expected frequency distributions were generated using exact binomial probabilities for each Reading passage at each grade it was tried out. These distributions compared

well with the empirical distributions, except at the extremes of the score distribution. Binomial distributions were then generated for the final sets of Mathematics and Spelling items. These data were used to estimate growth across grades on the items as a cluster, using the expected number of students attaining a possible cut-score as the measure of growth.

In closing, there are a few observations concerning the Rasch analysis that are important. The first approach we took for the Rasch analysis, that is, the vertical slice, was used because we wanted to generate item statistics separate by grade. Although we could explain reasonably well the inconsistencies we observed in the data, we were concerned that the small sample size may have been a factor. The second approach, that is, the horizontal slize, was used since it yielded the maximum number of students taking a set of items. This approach did not contain linking items, however, since the analysis was conducted by grade-referenced item clusters. There were several other possible approaches to the Rasch analysis, some of which we are planning to pursue. These approaches involve pairwise analyses in terms of both student grade and item cluster. For example, the two item clusters referenced to Grades 3 and 4 are common to both the Grade 3 and Grade 4 samples. The Grade 3/4 sample would contain approximately 200 students. After all similarly paired grade samples (i.e., Grades 1 and 2, Grades 2 and 3, Grades 3 and 4, etc.) are analyzed separately, the resulting distributions would then be equated. The apparent advantage of this type of approach is that it would increase both the range of item difficculties and the range of person abilities in each separate analysis while also maximizing the number of linking items and the sample size.

Table 1

BASIS Item Analysis Design

For Each Form of Each Test

Sample Size[1]

Students / Items	Grade Level 0	1	2	3	4	5	6	7	8	Total per items
8								100	200	300
7							100	100	200	400
6						100	100	100	200	500
5					100	100	100			300
4				100	100	100				300
3			100	100	100					300
2	100	100	100	100						400
1	100	100	100							300
Pre 1	100	100								200
Total Per Grade	100	100	100	100	100	100	100	100	200	

[1]Number of students per form per test.